Automatic Learning of the Morphology of Medical Language using Information Compression

Shamim Ara Mollah, MA and Stephen B. Johnson, PhD
Department of Biomedical Informatics, Columbia University, New York, NY

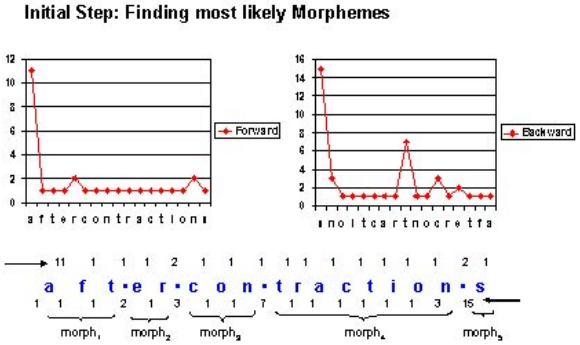
Conversion of free-text strings in a natural language to a standard representation (codes) is an important reoccurring problem in biomedical informatics. Determining the content of a string involves identifying its meaningful constituents (morphemes). One current method of identifying these constituents is to look them up in a preexisting table (lexicon). Manual construction of lexicons and grammars in complex domains such as biomedicine is extremely laborious. As an alternative to the lexico-grammatical approach, we introduce a segmentation algorithm that automatically learns lexical and structural preferences from corpora via information compression. The method is based on the Minimum Description Length (MDL) principle from classic information theory.

BACKGROUND

Word segmentation is a key problem in many natural language processing tasks. Typically, medical word segmentation is performed by morphological analysis based on lexical and grammatical knowledge. Morphological lexicons are domain specific, based on dictionary lookup or hand crafted rules. In medicine, these lexicons are very large. The size and complexity of these lexicon vocabularies contribute substantial problems for natural language processing. First, the medical lexicons contain compound medical terms; second, the vocabularies are not static[1]. The task of constructing them manually is laborious. The work presented in this poster offers a novel approach to this problem by employing an unsupervised learning algorithm within the Minimum Description Length (MDL) paradigm[2]. The fundamental idea behind the MDL principle is that any regularity in the data can be used to compress the data.

METHOD AND RESULTS

The task here is to find the optimal segmentation of the source text into morphemes. The cost function is derived from Minimum Description Length principle, which measures the goodness of the representation and the model complexity. Including a model complexity term generally improves generalization by inhibiting over learning. This is carried out in two steps: The first step takes words as inputs and gives an initial hypothesis regarding what stems and affixes are, based on an algorithm developed by Zellig Harris[4].



In the second step these morphemes are pruned based on the Mutual Information (MI) principle[3].

After Applying Mutual Information:
a f t e r • c o n • t r a c t i o n • s

We used 3.2 mb of UMLS terms as our training set. Our test set is composed of 1000 manually annotated words, which were chosen randomly from the UMLS. We achieved >80% sensitivity and >90% specificity.

CONCLUSIONS AND FUTURE WORK

As the size and the manual construction of the medical lexicons are becoming a growing concern, this technique can offer a plausible methodology for identifying meaningful components of language strings.
This method is also applicable for automatic discovery of hidden structure in other kinds of sequential data, particularly DNA sequences. In the future, we intend to refine our statistical methods and direct our focus towards identifying word boundaries in sentences.

REFERENCES

1. C. Lovis, PA. Michel, R. Baud, JR. Scherrer. Word segmentation processing: A way to exponentially extend medical dictionaries. MEDINFO 1995; 8 pt 1:28-32.
2. Jorma Rissanen. Stochastistic complexity in statistical inquiry. World Scientific Series in Computer Science, 15:79-93, 1989.
3. K Church and P. Hanks Word association norms, mutual information, and lexicography. Proceedings of ACL27, 76-83, 1989.
4. Zellig Harris. Morpheme boundaries within words: Report on a computer test. In transformations and Discourse Analysis. Papers, volume 73, 1967.